

DOCUMENT RESUME

ED 309 198

TM 013 687

AUTHOR van der Linden, Wim J.
TITLE The Use of Test Scores for Classification Decisions with Threshold Utility. Project Decision-Theoretic Aspects of Test Use No. 1. Research Report 86-1.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 86
NOTE 35p.
AVAILABLE FROM Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research/Technical (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Bayesian Statistics; *Classification; *Decision Making; Elementary School Students; Foreign Countries; Grade 6; Intermediate Grades; Regression (Statistics); *Scores; Statistical Analysis; *Test Use
IDENTIFIERS Linear Models; Netherlands; *Thresholds; *Utility Functions

ABSTRACT

Differences between traditional linear regression and a Bayesian approach to classification are discussed. Classification consists of assigning subjects to one of several available treatments on the basis of their test scores when the success of each treatment is measured by a different criterion. Formulating this problem as an empirical Bayes decision problem is indicated. As an example, the case of classification with a threshold utility function is analyzed and optimal assignment rules are derived. Data from over 15,500 students taking the achievement test of the Dutch National Institute of Educational Measurement at the end of grade 6, usually used to get an impression of the predictive validity of the test, were used as an example to explore the possibility of establishing optimal rules for the selection of continuation schools using this theory of classification. The following types of secondary education were selected as "treatments" in the analysis: (1) lower vocational education; (2) lower general education; and (3) middle general education. Students who continued at the same school or a higher one were considered successes. Findings were that differences between the two types of general education are very small. Three tables give data for the example study, and one graph illustrates the logistic regression lines for the three treatments. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED309198

The Use of Test Scores for Classification Decisions with Threshold Utility

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Research
Report
86-1

Wim J. van der Linden

TM013687

Department of Education

Division of
Educational Measurement
and Data Analysis

University
of
Twente

Project Decision-Theoretic Aspects of Test Use No. 1

Colophon

Typing : A. Burchartz

Cover design : M. Driessen, AV-section, University of Twente

Printed by : Central Reproduction Department, University of Twente

The Use of Test Scores for Classification
Decisions with Threshold Utility

Wim J. van der Linden

To appear in Journal of Educational Statistics

Abstract

The classification problem consists of the task to assign subjects to one out of several available treatments on the basis of their test scores where the success of each treatment is measured by a different criterion. It is indicated how this problem can be formulated as an (empirical) Bayes decision problem. As an example, the case of classification with a threshold utility function is analyzed and optimal assignment rules are derived. The results are illustrated empirically with data from a classification problem in which achievement test data are used to assign students to appropriate continuation-schools.

The Use of Test Scores for Classification
Decisions with Threshold Utility

One of the basic problems in test-based decision making is the classification problem. Typically, the classification problem confronts the decision maker with several "treatments" to which subjects have to be assigned. For each treatment there is a qualitatively different criterion representing the success of exposing the subject to the treatment. The only thing the decision maker has at his/her disposal is the subjects' scores on a certain aptitude or performance test administered prior to the treatment.

Examples of test-based classification problems include vocational guidance situations in which most promising types of schools or careers must be identified, classification of students in individualized instructional systems with tracks at different levels, testing for the military service, selecting one out of different available therapies in clinical settings, etc.

Obviously, the classification problem is a decision problem, and as such it has definite relationships to other decision problems in educational and psychological testing. Elsewhere the author has proposed a typology of test-based decision making which, in addition to the classification decision, has the selection, mastery, and placement decisions as basic types of decision making (van der Linden, 1985). The classification decision resembles the placement decision in that the latter is a treatment-assignment problem as well. The crucial difference, however, is that in

placement decisions each treatment leads to the same criterion whereas in classification decisions each treatment has a different criterion. An example of a placement decision is the aptitude-treatment-interaction paradigm in instructional psychology where students are allowed to reach the same educational objectives via different instructional treatments. Selection and mastery decisions differ from classification and placement decisions by the presence of only one treatment. In these decisions, it is the decision maker's task to decide whether or not to accept subjects for a certain treatment, respectively, whether or not they have profited enough from a treatment to be dismissed. The four basic types of decisions briefly defined here can be met both in their pure forms or in combinations with each other. The latter is the case, for instance, in decision making in individualized study systems which can be conceived of as networks consisting of these various types of decisions as nodes (van der Linden & Vos, 1986). Also, further refinements within each type are possible, for instance, by imposing quota restrictions or distinguishing between subpopulations varying on a relevant attribute (cf. van der Linden, 1985).

An appropriate framework for dealing with decision problems such as the above is (empirical) Bayesian decision theory (e.g., Raiffa, 1968). Applications of Bayesian decision theory to selection (e.g., Chuang, Chen, & Novick, 1981; Cronbach & Gleser, 1965; Novick & Lindley, 1979), mastery (e.g., Hambleton & Novick, 1973; Huynh, 1976; van der Linden, 1980; van der Linden & Mellenbergh, 1977),

and placement decision problems (e.g., Cronbach & Gleser, 1965; van der Linden, 1981) are amply available now, whereas Bayesian theory has also been used to deal with the problem of selection from different subpopulations (e.g., Mellenbergh & van der Linden, 1981; Petersen, 1976; Petersen & Novick, 1976). To date, however, the classification problem has been devoid of a decision-theoretic analysis. The traditional approach to this type of problem has been the use of linear-regression techniques in which for each treatment the regression line of the utility of its criterion scores on the test scores is estimated and subjects are assigned to treatments on the basis of their predicted utility scores. In the case of a test battery instead of a single test, multiple-regression analysis has been the choice. This approach has led to the concept of differential validity (Horst, 1955; Mollenkopf, 1950) which has had a prominent place in the literature on constructing tests for classification decisions. The difference between the above approach and a decision-theoretic one is that the former first models the problem as a (continuous) point prediction problem and only then introduces the aspect of classifying subjects into different treatments on the basis of their predicted utility scores. It will become clear below that a decision-theoretic approach, on the other hand, directly models the problem as a decision problem with discrete actions. This provides the decision maker with the possibility of weighting decision errors explicitly via the specification of a utility function. Also, the model takes into account the full (posterior) distributions of criterion scores

given the test score, and, consequently, does not necessarily entail such assumptions as linear regression of criterion on predictor scores and homoscedasticity of "prediction errors". However, specific conditions on the probability distributions and utility functions exist for which the Bayesian solutions boil down to the linear regression solutions. An example of this will be given below.

The purpose of this paper is to demonstrate how the classification problem can be formalized as a problem of Bayesian decision making. In particular the case of classification with a threshold utility function is analyzed, and for this case it is indicated how, for a variety of conditions, optimal rules can be found. As an example of classification decisions with threshold utility a problem is considered that is well known in certain guidance situations, namely the choice of an appropriate continuation-school. A case with empirical data will be used to illustrate the selection of an optimal decision rule for this problem.

Before embarking on this, it is re-emphasized that, though much of the logic in the following resembles that of the Bayesian approach to the mastery, selection, and placement problems dealt with in the papers earlier referred to, the classification problem has the distinctive feature of possessing more than one treatments each having its own criterion. As a consequence, the probability distributions and the utility functions in the model are all indexed by a different treatment and defined on a different

criterion. In a more formal sense, it is exactly the goal of this paper to explore the consequences of this treatment dependency for the Bayesian approach.

Classification Decisions with Threshold Utility

Suppose a series of subjects, who can be considered as randomly drawn from some population P , must be classified into $t + 1$ treatments indexed by $j = 0, 1, \dots, t$. Each treatment leads to a certain distribution of performances of P on its associated criterion which is denoted by a random variable Y_j with range R_j . The variable Y_j will be considered to be continuous (although in many applications it may be discrete). The test scores observed prior to the treatment are denoted by a random variable X with discrete values $x = 0, \dots, n$. It is assumed that P yields a joint distribution of test and criterion scores with probability (density) function $\eta_j(x, y_j)$. The classification problem consists of the task to define a decision rule on x such that the expected utility with respect to sampling from P is maximal.

In the present paper a somewhat more specific problem is considered for which it can be assumed that the optimal decision rule has the monotone form of a series of cutting scores on the test:

$$(1) \quad 0 = c_0 \leq c_1 \leq \dots \leq c_j \leq \dots \leq c_{t+1} = n \quad \text{with } t \leq n,$$

where treatment j is assigned in the event of $c_j \leq X < c_{j+1}$ (the second inequality not being strict for $j = t$).

Conditions for monotone Bayes actions with respect to a single parameter are given, for example, in Chuang, Chen, and Novick (1981), Ferguson (1967, sect. 6.1), and Lindgren (1976, sect. 8.3.5), and go back to a basic result by Karlin and Rubin (1956). In general they imply that the posterior distribution of the parameter is stochastically increasing and that there is an ordering of the actions for which the difference in utility between adjacent pairs of actions changes sign at most once. For the classification problem as formalized above this does not hold without modifications. A set of conditions sufficient for monotone solutions for classification with threshold utilities will be given below.

As an example of a classification decision, consider the case of selecting an optimal continuation-school as this can be met in testing for vocational-guidance purposes. It is assumed that each of the available schools teaches globally the same subject matter but at different levels so that it is possible to order them with respect to their utilities. Also, each school is finished with a certifying exam which the student may pass or fail. For such cases it seems obvious to define utilities that vary with the levels of attainments at the end of the schools. However, if only the dichotomy pass-fail is observed, threshold utilities may be a proper choice. The following threshold utility function is defined:

$$(2) \quad u_j(y_j) = \begin{cases} w_j & \text{for } y_j \geq d_j \\ v_j & \text{for } y_j < d_j, \end{cases}$$

with

$$w_j > v_j \quad \text{for } j = 0, \dots, t,$$

where y_j can be interpreted as, e.g., the grade-point average at the end of school j , d_j is the minimal value of y_j for which a student passes, and w_j and v_j are the utility of finishing school j successfully and not successfully, respectively. In addition to (2), it is assumed that the following conditions holds:

$$(3) \quad w_{j-1} - v_{j-1} \leq w_j - v_j, \quad j = 1, \dots, t,$$

$$(4) \quad Q_j(d_j|x) \text{ is decreasing in } x, \quad j = 0, \dots, t,$$

$$(5) \quad Q_{j-1}(d_{j-1}|x) - Q_j(d_j|x) \text{ is increasing in } x, \quad j = 1, \dots, t,$$

where $Q_j(\cdot|x)$ is defined as the distribution function of Y_j given $X = x$. Since in (2) $w_j > v_j$ for $j = 0, \dots, t$, it follows that (3) is equivalent to the condition of monotone utility in the Bayesian decision problem with a single parameter (e.g., Ferguson, 1967, p. 285). Observe that (3) orders the treatments with respect to $w_j - v_j$. The condition in (4) corresponds with the condition of a

stochastically increasing posterior distribution in the standard problem of a monotone Bayes rule with respect to a single parameter (Chuang, Chen, & Novick, 1981, theorem 2), but is now required to hold only for $Y_j = d_j$. Finally, the condition in (5) relates the posterior distributions to each other stating that for the treatments ordered higher by their utility difference the probability of a failure decreases quicker in x . Whether this condition holds is dependent on the difficulty of the test for the students in the population. For a test of moderate difficulty the probabilities of a failure at the lower end of the test score scale will be ordered such that a lower treatment has a smaller probability. Hence, the lower the value of j , the less space there is for the probability to decrease in x , and (5) is likely to hold. If the test gets more difficult, the situation may be more complicated. In each new application it must be checked if (4) - (5) hold.

It follows from the derivations below that (3) - (5) are sufficient for a monotone solution to a classification problem with threshold utility function (2).

As monotone solutions are looked for, the expected utility when sampling from P can be written as

$$(6) \quad B(c_1, \dots, c_t) \equiv \sum_{j=0}^t \sum_{x=c_j}^{c_{j+1}-1} \int_{R_j} u_j(y_j) \eta_j(x, y_j) dy_j.$$

The set of optimal cutting scores in the Bayesian sense is the choice of values for $(c_1, \dots, c_t)$ maximizing (6).

Case of $t = 1$

First the case of two treatments ($t = 1$) is considered and then the results will be generalized to larger numbers of treatments ($t > 1$).

Substituting (2) into (6) for $t = 1$ and omitting the index of c gives

$$(7) \quad B(c) = \sum_{x=0}^{c-1} \int_0^{d_0} v_0 \eta_0(x, y_0) dy_0 + \sum_{x=0}^{c-1} \int_{d_0} w_0 \eta_0(x, y_0) dy_0 +$$

$$\sum_{x=c}^n \int_0^1 v_1 \eta_1(x, y_1) dy_1 + \sum_{x=c}^n \int_{d_1} w_1 \eta_1(x, y_1) dy_1.$$

Using

$$\int_0^{d_j} \eta_j(x, y_j) dy_j = Q_j(d_j | x) \lambda(x),$$

where

$$\lambda(x) = \lambda_j(x), \quad j = 0, \dots, t,$$

is the marginal probability function of X , completion of the first two sums gives

$$(8) \quad B(c) = \{ \text{constant} - \sum_{x=c}^n v_0 q_0(d_0|x) \lambda(x) \} + \{ \text{constant} -$$

$$\sum_{x=c}^n w_0 [1 - q_0(d_0|x)] \lambda(x) \} + \sum_{x=c}^n v_1 q_1(d_1|x) \lambda(x) \\ + \sum_{x=c}^n w_1 [1 - q_1(d_1|x)] \lambda(x).$$

This can be reduced further to

$$(9) \quad B(c) = \text{constant} + \sum_{x=c}^n \{ (w_0 - v_0) q_0(d_0|x) - (w_1 - v_1) q_1(d_1|x) + w_1 - w_0 \} \lambda(x).$$

The assumption in (5) states that $q_0(d_0|x) - q_1(d_1|x)$ is increasing in x . Thus, if $x_1 < x_2$, it holds that $q_0(d_0|x_1) - q_1(d_1|x_1) \leq q_0(d_0|x_2) - q_1(d_1|x_2)$, or $q_0(d_0|x_1) - q_0(d_0|x_2) \leq q_1(d_1|x_1) - q_1(d_1|x_2)$. From (4) it follows that $q_j(d_j|x_1) - q_j(d_j|x_2)$ is nonnegative while (2) - (3) imply that $0 < w_0 - v_0 \leq w_1 - v_1$. Hence, it applies that $(w_0 - v_0)[q_0(d_0|x_1) - q_0(d_0|x_2)] \leq (w_1 - v_1)[q_1(d_1|x_1) - q_1(d_1|x_2)]$. But then $(w_0 - v_0)q_0(d_0|x_1) - (w_1 - v_1)q_1(d_1|x_1) \leq (w_0 - v_0)q_0(d_0|x_2) - (w_1 - v_1)q_1(d_1|x_2)$ and the bracketed quantity in (9) is increasing in x as well. Since $\lambda(x)$ is nonnegative for all values of x , it thus holds that (9) is maximal for the smallest value of x for which

$$(10) \quad (w_0 - v_0)q_0(d_0|x) - (w_1 - v_1)q_1(d_1|x) + w_1 - w_0$$

is positive. This solution is unique if in the region of x values in which the expression in (9) changes sign $\lambda(x)$ is nonzero and the monotonicity condition in (5) is strict (and so for at least one treatment (4) is strict as well). If one of these conditions is not met, more than one solution may exist which are all equivalent in the sense of yielding the same expected utility. In practice, the solution is found by estimating the conditional distributions of criterion scores given predictor scores, checking for (5), and substituting the estimates into (10). It is also possible to fit a model for the distributions and estimate the solution under this model. Small-sample solutions of (10) can be obtained only for models with known parameters.

Case of $t > 1$

The generalization to more than two treatments with threshold utility follows from a derivation along the same line as in (6) - (10). The results turn out to be simple and amount to applying the above findings to each pair of adjacent treatments. In principle, this procedure parallels the one for placement decisions with more than two treatments (van der Linden, 1981).

A Linear Regression Solution

As noted earlier, specific conditions on the probability distributions and utility functions exist for which the Bayes solution boils down to a linear regression one. This will be shown

now.

As the linear regression approach is not concerned with any differences in utilities between treatments, it is assumed that w_j and v_j in (2) take the same values for all treatments. It follows that $w_0 - v_0 = w_1 - v_1$ and $w_1 - w_0 = 0$, and that (10) reduces to

$$(11) \quad Q_0(d_0|x) - Q_1(d_1|x).$$

If normal distributions of Y_j given $X = x$ can be assumed, (11) is replaced by

$$\frac{d_0 - E_0(Y|x)}{\text{Var}_0(Y|x)^{\frac{1}{2}}} - \frac{d_1 - E_1(Y|x)}{\text{Var}_1(Y|x)^{\frac{1}{2}}},$$

where $E_j(Y|x)$ and $\text{Var}_j(Y|x)$ are the conditional expectations and variances, respectively. The choice of normal distributions can be defended from a curve-fitting point of view. It should then be checked whether this assumption is reasonable for the data at hand and whether conditions (4) - (5) are met satisfactorily.

Further approximations are possible. For example, if it holds (approximately) that $\text{Var}_0(Y|x) = \text{Var}_1(Y|x)$ for relevant values of x , then the optimal cutting score between the two treatments is the smallest value of x for which

$$(12) \quad E_1(Y|x) - E_0(Y|x) > d_1 - d_0$$

holds. Also, for linear regression lines $E_j(Y|x) = \alpha_j + \beta_j x$ and homoscedasticity, it follows that $\text{Var}_j(Y|x) = \text{Var}_j(Y.x)$. The optimal cutting score c^* is now equal to the smallest integer value larger than or equal to

$$(13) \quad \frac{(d_1 - \alpha_1) \text{Var}_0(Y.x)^{\frac{1}{2}} - (d_0 - \alpha_0) \text{Var}_1(Y.x)^{\frac{1}{2}}}{\beta_1 \text{Var}_0(Y.x)^{\frac{1}{2}} - \beta_0 \text{Var}_1(Y.x)^{\frac{1}{2}}}$$

To calculate this quantity only estimates of the regression parameters and the pooled variances of Y given $X = x$ are required.

Although the distributional assumptions in (13) refer to the regression lines of criterion on predictor scores, this solution still does not coincide with the results of the traditional linear regression approach. The difference is the presence of the thresholds d_j and the pooled variances $\text{Var}_j(Y.x)$ in (13). As soon as d_j and $\text{Var}_j(Y.x)$ are treatment-independent constants, (13) reduces further to

$$(14) \quad (\alpha_0 - \alpha_1) / (\beta_1 - \beta_0)$$

which is the value of x for which the regression lines cross and on which treatment assignment in the traditional approach is based.

As an alternative to the assumptions leading to (12) - (14), it

is also possible to model (11) directly and assume, for instance, a normal or logistic distribution function as a model for the probabilities of failure as a function of x . If the choice is the logistic function, which, as is well known, leads to results hardly discriminable from the normal but offers computational advantages, the analysis in (11) boils down to logistic regression analyses for the two treatments, and the optimal cutting score is found as the value of x for which the logistic regression lines cross. If such values occur outside the range of test scores, there is a treatment dominating the others and the assignment of all subjects to this treatment is the solution. Again, it must be checked if these logistic curves fit the data. If they do for a negative value of the slope parameter, then (4) is automatically met; but it must always be checked separately if (5) holds for the range of the data. The literature shows successful examples of fitting logistic and normal functions to test data. For example, Raju, Burke, Normand, and Dye (1984) followed logistic regression analysis for a selection problem. Although their problem had only one treatment, the fact that they found a good fit of the logistic model suggest a successful application of logistic regression analysis to classification data. An analysis comparable to the one by Raju et al., with the normal instead of the logistic distribution function, is given in Dagenais (1984).

An Empirical Example

The example in this section gives an empirical illustration of the problem of choosing an appropriate continuation-school.

In The Netherlands primary education ends at grade six and is followed by a complicated system of secondary education with types of schools providing education from lower level vocational to university-track programs. By law it is required that the transition from primary to secondary education is partly based on empirical evidence about the pupil's suitability obtained from, for instance, the administration of an achievement test or a psychological test battery. An achievement test popular for this purpose is the Eindtoets basisonderwijs prepared annually by the Dutch National Institute of Educational Measurement (Cito). Periodically, Cito gathers flow data on pupils relating their test scores to what happened to them one year later. These data are only used to get an impression of the predictive validity of the test, not to base decision rules for school selection on.

In the analyses reported here data from a 1981 Cito study of a certain regional unit in The Netherlands were used to explore the possibility of establishing optimal rules for school selection using the theory of classification decisions with threshold utility in this paper. As a model for the probability of success as a function of the achievement test score the logistic function was chosen. A pupil who passed his/her first year and remained in the same type of school or left it for a higher level school was

considered a success; all other cases were considered a failure. The following types of secondary education were selected as "treatments" in the analyses: Lower Vocational Education (LVE), Lower General Education (LGE), and Middle General Education (MGE). Table 1 gives the empirical proportions of successes and the

Insert Table 1 about here

number of cases for each treatment. As only grouped data were available, logit analysis of the proportions was applied for the middles of the test score intervals reported in the table. Some of the data in the outmost intervals had to be pooled to bring the number of cases to a level acceptable for a likelihood-ratio goodness-of-fit test. The bottom line of the table reports the probabilities of exceedance of the chi-square values and shows that the logit model had a satisfactory fit to the data. Table 1 also gives the slopes and intercepts of the logistic regression lines for each treatment. The numbers of cases in Table 1 justify the large-sample approach adopted in this illustration. The empirical proportions of success in Table 1 suggest that the monotonicity conditions in (4) - (5) also hold for the range of individual test scores in the population. As is clear from Figure 1 below, the condition in (4) is met. For LVE/LGE the condition in (5) applies for the whole range of test scores. The same holds for LGE/MGE with a minor exception at the lower end of the range of test scores

($x \leq 4$) where the difference in (5) decreases instead of increases.

In the following analyses it was first assumed that the utilities were the same for each treatment, and then the effects of deviations from the equalities $w_j - v_j = w_{j-1} - v_{j-1}$ and $w_j = w_{j-1}$ on the behavior of the optimal cutting scores were studied. If the utilities are the same for each treatment, the solution is given by (11) and is found as the smallest value of x not to the left of the point at which the logistic regression lines cross. Figure 1 shows the regression lines for each type of program in the example. It appears that the optimal point for deciding between LVE and LGE is

Insert Figure 1 about here

$x = 47$ (the lines cross at $x = 46.9$), whereas the same point for LGE and MGE falls at the border of the range of test scores (the lines cross slightly over $x = 50$ implying that for a longer test the cutting score would be at $x = 51$). The general impression from this result is that for almost all possible test scores the choice of LVE is optimal; only for very high test scores does the choice of LGE appear to be best.

The aspect of Figure 1 most conspicuous, however, is the near coincidence of the lines for LGE and MGE. Although LGE is a better choice for the whole range of test scores, it hardly dominates MGE. On the other hand, there is much difference between these two

general types of education and LVE. These properties explain the results in Table 2 where it is displayed how the cutting scores

Insert Table 2 about here

behave as a function of departures from the case with equal values for $w_j - v_j$. The optimal cutting score between LVE and LGE appears to be robust in the neighborhood of the utility ratio equal to one. A change of 20 percent in the utility difference for LGE relative to LVE results only in a change of 3 points for the cutting score. The decision between MGE and LGE, however, depends heavily on the utility ratio. Due to the fact that the regression lines nearly coincide, a negligible decrease of the utility ratio immediately makes the cutting score go to the lower end of the test score scale.

It should be noted that in Table 2 it is still assumed that $w_j - w_{j-1} = 0$. The opposite case, in which $w_j - w_{j-1}$ deviates from zero but $w_j - v_j = 1$, is presented in Table 3. It appears that the

Insert Table 3 about here

optimal cutting score between LGE and MGE shows exactly the same all-or-none behavior as in Table 2. The optimal cutting score between

LVE and LGE varies more regularly as a function of $w_j - w_{j-1}$. It should be noted that the optimal cutting score between LVE and LGE decreases in $w_j - w_{j-1}$. This is due to the assumptions underlying Table 3. As $w_j - v_j = 1$, an increase in w_j relative to w_{j-1} also implies a relative increase of v_j . So treatment j becomes increasingly attractive compared to $j-1$ and the optimal cutting score decreases. Table 3 shows that on the whole the optimal cutting score is more sensitive to changes in the neighborhood of $w_j - w_{j-1} = 0$ than for larger values of w_j relative to w_{j-1} . This is a favorable result because among students and parents a LGE diploma is generally appreciated much better than a LVE diploma.

No doubt the most important result from this empirical example is that the choice between LGE and MGE is mainly a matter of subtle differences in utilities. In the relevant range of test scores, differences between the probabilities of success are hardly discernible. These findings seem to lend support to the belief held by some in The Netherlands that the differences between these two types of general education are too small to justify their separate existencies.

Concluding Remarks

Two further comments are necessary. First, it is noted that the decision maker's task of specifying his/her utility parameter values, usually a delicate matter, is a somewhat simpler one in the present case of a threshold utility function. The decision maker

only needs to provide values for the parameters w_j and v_j . In practical applications, this task can easily be performed using a lottery method well-known in this area (e.g. Luce & Raiffa, 1957, chap. 2) or via direct scaling of preferences (Vrijhof, Mellenbergh, & van den Brink, 1983). If it is believed that the final exam is not the ultimate criterion but that this lies in its civil effects, psychological well being, etc., Keeney and Raiffa's handling of proxy attributes may be useful (Keeney & Raiffa, 1976, sect. 2.5). The robustness analyses in Tables 2-3 show, however, that the optimal cutting scores may be sensitive to changes in the utility parameters for some ranges of possible values. Therefore, in general care should be taken when specifying their values.

The foregoing also showed the conditions under which the Bayesian approach yields the same solution as the linear-regression approach. The linear-regression solution can be conceived of as a Bayes rule under threshold utility (2) with equal parameter values for all treatments, homoscedasticity of the variances of Y given $X = x$ for all treatments, and equality of these conditional variances across all treatments. It seems unlikely that these restrictions will often be met in practice.

It was the purpose of this paper to give the reader the flavor of such differences between the traditional linear regression and a Bayesian approach to the classification problem.

References

- Chuang, D.T., Chen, J.J., & Novick, M.R. (1981), Theory and practice for the use of cut-scores for personnel decisions. Journal of Educational Statistics, 6, 129-152.
- Cronbach, L.J., & Gleser, G.L. (1965), Psychological tests and personnel decisions (2nd ed.). Urbana, Ill.: University of Illinois Press.
- Dagenais, D.L. (1984), The use of a probit model for the validation of selection procedures. Educational and Psychological Measurement, 44, 629-645.
- Ferguson, T.S. (1967), Mathematical statistics: A decision theoretic approach. New York: Academic Press.
- Horst, P. (1955), A technique for the development of a multiple absolute prediction battery. Psychological Monograph, No. 390.
- Hambleton, R.K., & Novick, M.R. (1973), Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.
- Huynh, H. (1976), Statistical considerations of mastery scores. Psychometrika, 41, 65-79.
- Karlin, S., & Rubin, H. (1956), The theory of decision procedures for distributions with monotone likelihood ratio. Annals of Mathematical Statistics, 27, 272-299.
- Keeney, R.L., & Raiffa, H. (1976), Decisions with multiple objectives: Preferences and value trade offs. New York: Wiley.

- Luce, R.D., & Raiffa, H. (1957), Games and decisions. New York: John Wiley & Sons, Inc., 1957.
- Mellenbergh, G.J., & van der Linden, W.J. (1981), The linear utility model for optimal selection. Psychometrika, 46, 283-293.
- Mollenkopf, W.G. (1950), Predicted differences and differences between predictions. Psychometrika, 15, 409-417.
- Novick, M.R., & Lindley, D.V. (1978), The use of more utility functions in educational applications. Journal of Educational Measurement, 15, 181-191.
- Petersen, N.S. (1976), An expected utility model for "optimal" selection. Journal of Educational Statistics, 1, 333-358.
- Petersen, N.S., & Novick, M.R. (1976), An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.
- Raiffa, H. (1968), Decision analysis. Reading, Mass.: Addison-Wesley.
- Raju, N.S., Burke, M.J., Normand, J., & Dye, D.A. (1984), The invariance of job parameters in a latent trait model for personnel selection. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, April 23-27.
- van der Linden, W.J. (1980), Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 4, 469-492.

- van der Linden, W.J. (1981), Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. Psychometrika, 46, 257-274.
- van der Linden, W.J. (1985), Decision theory in educational research and testing. In T. Husen & T.N. Postlethwaite, International encyclopedia of education: Research and studies. Oxford: Pergamon Press.
- van der Linden, W.J., & Mellenbergh, G.J. (1977), Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1, 593-599.
- van der Linden, W.J., & Vos, H.J. (1986), Optimale regels voor toetsgebruik in individuele studiesystemen Optimal rules for test use in individualized study systems . In W.J. van der Linden (Ed.), Moderne methoden voor toetsconstructie en -gebruik. Lisse, The Netherlands: Swets & Zeitlinger.
- Vrijhof, B.J., Mellenbergh, G.J., & van den Brink, W.P. (1983), Assessing and studying utility functions in psychometric decision theory. Applied Psychological Measurement, 7, 341-357.

Author Notes

The author is indebted to Michel A. Zwarts for providing the data for the empirical example.

Request for reprints should be sent to Wim J. van der Linden, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, Postbus 217, 7500 AE Enschede, The Netherlands.

Table 1
Empirical Proportion of Successes as a Function
of Test Scores for the Three Treatments

Test Score	Proportion of Successes After 1 Yr.		
	LVE	LGE	MGE
0 - 5			
6 - 10	0.897	0.575	
11 - 15			0.571
16 - 20	0.929	0.619	
21 - 25	0.947	0.760	
26 - 30	0.948	0.840	0.788
31 - 35	0.952	0.890	0.860
36 - 40	0.959	0.930	0.920
41 - 45		0.960	0.960
46 - 50	0.979	0.960	0.988
No. of Cases	1333	15926	2298
Slope	0.031	0.095	0.099
Intercept	2.2	-0.8	-1.0
Model Fit	0.641	0.071	0.105

Table 2
Optimal Cutting Scores Between the Treatments
as a Function of the Utility Ratio

Utility Ratio	Optimal Cutting Score	
	LVE/LGE	LGE/MGE
0.10	-	-
0.20	17	-
0.30	27	-
0.40	32	-
0.50	36	-
0.60	39	-
0.70	42	-
0.80	44	-
0.90	46	-
1.00	47	-
1.10	49	-
1.20	50	-

Note Utility ratio is defined as $(w_j - v_j)(w_{j-1} - v_{j-1})^{-1}$. It is assumed that $w_j = w_{j-1}$. Dash indicates cutting score outside range of test scores.

Table 3
Optimal Cutting Scores Between the Treatments
as a Function of $w_j - w_{j-1}$

$w_j - w_{j-1}$	Optimal Cutting Score	
	LVE/LGE	LGE/MGE
-0.10	-	-
0.00	47	-
0.10	28	-
0.20	20	-
0.30	15	-
0.40	10	-
0.50	5	-
0.60	-	-

Note It is assumed that $w_j - v_j = w_{j-1} - v_{j-1} = 1$.

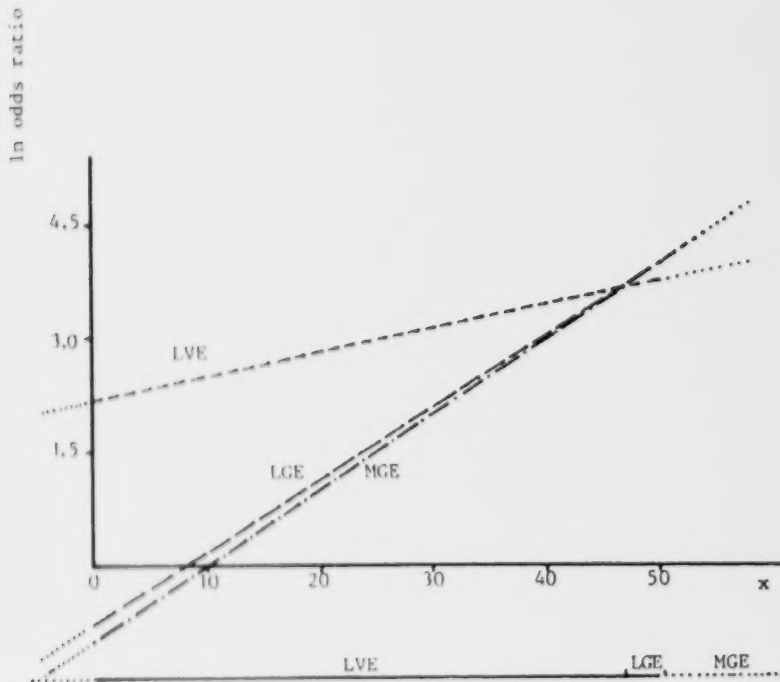
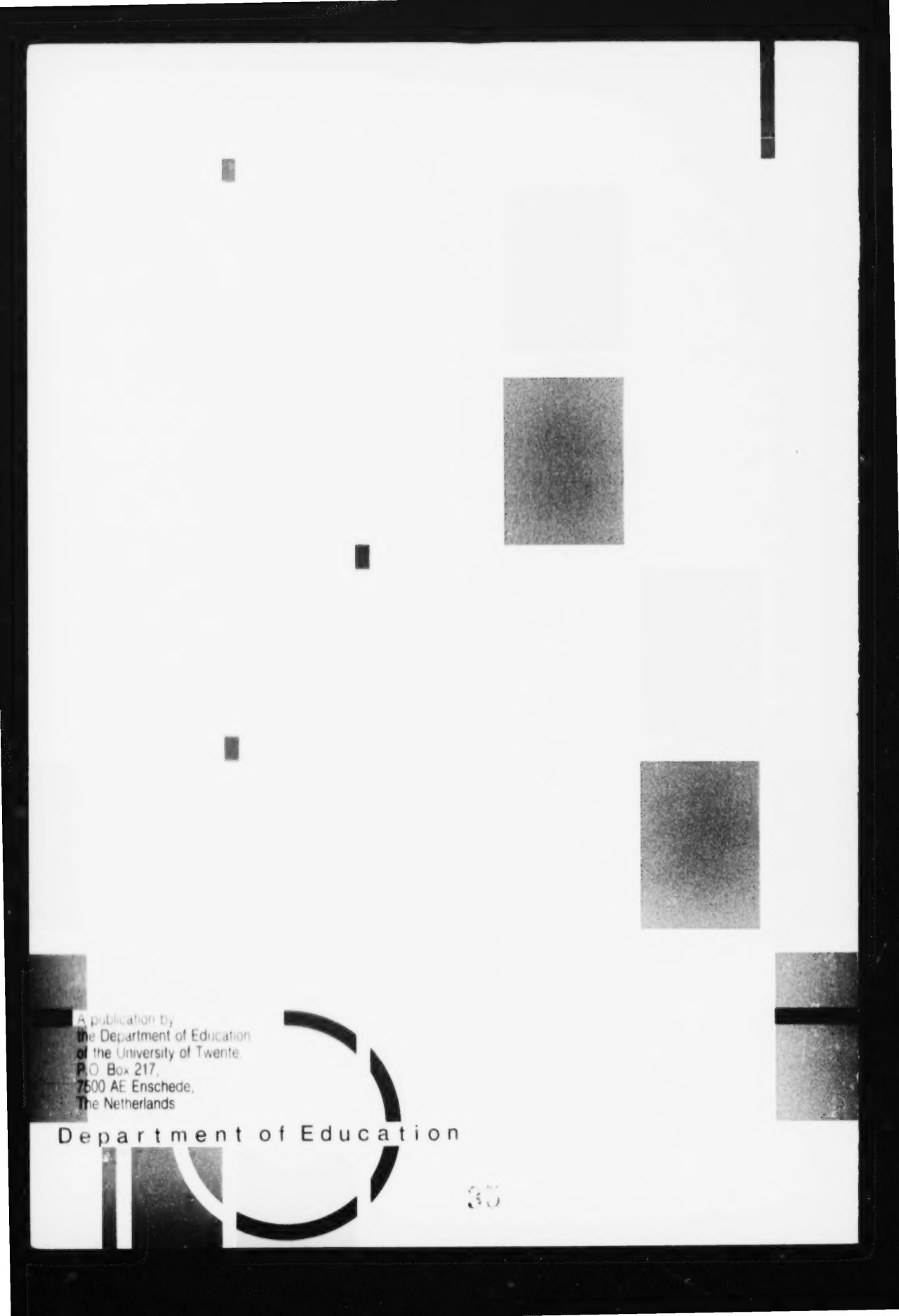


Figure 1. Logistic regression lines for the three treatments.

Titles of Recent Research Reports

- RR-86-1 W.J. van der Linden, The use of test scores for
classification decisions with threshold utility
- RR-86-2 H. Kelderman, Item bias detection using the loglinear
Rasch model: Observed and unobserved subgroups
- RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by
zero-one programming
- RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one
programming approach to Gulliksen's random matched
subtests method
- RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity
analysis with k sets of variables: An alternating least
squares method with optimal scaling features
- RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes
approach to item banking
- RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of
parallel tests by zero-one programming
- RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models
for paired comparisons with ties



A publication by
the Department of Education
of the University of Twente,
P.O. Box 217,
7500 AE Enschede,
The Netherlands

Department of Education